J. Appl. Prob. **52,** 269–277 (2015)
Printed in England
© *Applied Probability Trust* 2015

A QUEUEING LOSS MODEL WITH HETEROGENEOUS SKILL BASED SERVERS UNDER IDLE TIME ORDERING POLICIES

BABAK HAJI [*] [**] AND
SHELDON M. ROSS,[*] [***] *University of Southern California*

Abstract

We consider a queueing loss system with heterogeneous skill based servers with arbitrary service distributions. We assume Poisson arrivals, with each arrival having a vector indicating which of the servers are eligible to serve it. An arrival can only be assigned to a server that is both idle and eligible. Assuming exchangeable eligibility vectors and an idle time ordering assignment policy, the limiting distribution of the system is derived. It is shown that the limiting probabilities of the set of idle servers depend on the service time distributions only through their means. Moreover, conditional on the set of idle servers, the remaining service times of the busy servers are independent and have their respective equilibrium service distributions.

Keywords: Heterogeneous server; queueing loss system; limiting probability; no-memory policy; method of stages; reverse chain; equilibrium distribution; Gibbs sampler

2010 Mathematics Subject Classification: Primary 60K25
Secondary 60J27; 90B22

1. Introduction

The model we consider in this paper supposes that arrivals come to an n server system in accordance with a Poisson process with rate λ, and that each arrival has a vector of binary values $(x_1, \dots, x_n)$, with the interpretation that server i is eligible to serve that arrival if $x_i = 1$ and is ineligible if $x_i = 0$, $i = 1, \dots, n$. The binary vectors of successive arrivals are assumed to be independent and identically distributed having a specified distribution. An arrival can be assigned to any of the servers that are both currently idle and eligible to serve that arrival; if there are no such servers, the arrival is lost. The time it takes server i to serve a customer has a general distribution G_i, $i = 1, \dots, n$.

The preceding model, under the assumption that the assigning rule used is to assign an arrival to the idle/eligible server that has been idle the longest since its last service completion, was introduced by Adan and Weiss [1]. (Although Adan and Weiss used different terminology by classifying arrivals according to their eligibility vectors, the models are mathematically equivalent.) Using a supplementary variable approach to make their system Markovian, they derived the limiting distribution for this model, and in doing so showed that the limiting

Received 12 June 2013; revision received 22 February 2014.

[*] Postal address: Department of Industrial and Systems Engineering, University of Southern California, Los Angeles, CA 90089, USA.

This material is based upon work supported by the National Science Foundation under contract/grant number CMMI1233337.

[**] Email address: bhaji@usc.edu
[***] Email address: smross@usc.edu

https://doi.org/10.1239/jap/1429282621 Published online by Cambridge University Press

distribution of the ordered set of idle servers depends on the service distributions only through their means.

In this paper we assume that a random eligibility vector $(X_1, \dots, X_n)$ is exchangeable, meaning that the joint probability mass function of $X_{i_1}, \dots, X_{i_n}$ is the same for all permutations $i_1, \dots, i_n$ of $1, \dots, n$. Because the X_i are binary, this is equivalent to the statement that for any $k \le n$, conditional on $\sum_{i=1}^n X_i = k$, all $\binom{n}{k}$ possible values of the vector $(X_1, \dots, X_n)$ are equally likely. (Although the distribution of $\sum_{i=1}^n X_i$ can be arbitrary, the most common case of exchangeability is when it is binomial, and so $X_1, \dots, X_n$ are independent and identically distributed Bernoulli random variables.) Although our exchangeability assumption is restrictive, we do allow for a general class of operating policies. Letting the 'idle servers vector' be 0 if there are currently no idle servers, or $i_1, \dots, i_k$ if there are currently k idle servers, with i_1 having been idle the longest, i_2 the second longest, and so on, we define an idle time ordering rule as one whose assignment decisions are based solely on the number of idle servers and the positions in the idle servers vector of those servers that are eligible to be assigned. Examples of such policies would be to assign to a randomly chosen idle and eligible server, or to assign to the idle eligible server that has been idle the longest, or that has been idle the shortest. Our analysis uses the *method of stages*, which starts by assuming that the service distributions are all general Erlang. Doing so enables us to analyze the model as a continuous-time Markov chain. Using a conjecture concerning the reverse chain enables us to find, up to a multiplicative constant, the limiting probabilities for this model, which surprisingly are the same no matter which idle time ordering policy is employed. We show that the limiting distribution of the vector for the idle servers depends on the service distributions only through their means, and that given the set of idle servers (a) all possible idle server vectors are equally likely, and (b) the remaining service times (as well as the amounts of service time each has so far provided) of the busy servers are independent and distributed according to their respective equilibrium service distributions. Application of a continuity argument then establishes these results for arbitrary service distributions. Because the determination of the multiplicative constant by summing all the probabilities is computationally intractable for large n, we show how the Gibbs sampler can be used to simulate a Markov chain whose stationary probabilities enable us to determine the desired quantities of interest for our model.

In Section 2 we introduce the model and provide some further notation. In Section 3 we review general Erlang distributions and the method of stages. In Section 4 we derive the stationary probabilities, and in Section 5 we present the Gibbs sampler simulation approach.

There have been a variety of papers whose authors have analyzed queueing loss models that allow an arrival to go to any idle server. For instance, Fakinos [4] studied the equilibrium behavior of an M/G/k loss system with heterogeneous servers under the policy that assigns arrival customers to the idle servers uniformly at random. By utilizing the method of supplementary variables, he gave a generalization of the Erlang B-formula and showed that the Erlang B-formula holds under the assumption of heterogeneous servers with equal mean service times. In another paper Fakinos [5] considered the same model but this time with balking customers. That is, customers will immediately depart the system upon their arrival with a certain probability, otherwise they will be served by one of the idle servers at random. Assuming equal means for the service time distributions, Fakinos showed that the Erlang B-formula is valid for the new model and proved that in equilibrium the customer (served, balking, or rejected) departures from the system form a Poisson process. Cooper and Palakurthi [3] studied a loss system with Poisson arrivals and heterogeneous servers with general service distributions. Assuming a priority ordering rule, which is a fixed permutation of the servers

https://doi.org/10.1239/jap/1429282621 Published online by Cambridge University Press

which is employed by always choosing the idle server that is earliest on that list, Cooper and Palakurthi showed by a counterexample that the loss probability does not depend on the service time distributions only through their means.

There have also been a variety of papers concerned with finding the policy that minimizes the rate of lost arrivals. One recent paper by Ross [9] which, like our model, considers an n server loss system in which each arrival comes with a realization of an exchangeable eligibility vector that indicates which of the idle servers it can use. Letting I_i be the indicator variable for whether arrival i is served or not, Ross showed, under the assumption of heterogeneous exponential service distributions, that for any arrival process that is independent of the service times and any value of r, the vector $(I_1, \dots, I_r)$ is stochastically maximized when each arrival is assigned to the idle eligible server with the largest service rate. Other references concerning the optimization problem can be found in [7].

It has also recently been brought to our attention that results having some similarity to ours have been obtained by Gopalakrishnan *et al.* [6]. The model in [6] is similar but less general than our model in that it assumes exponential service distributions and that all servers are always eligible (that is, $\mathbb{P}(X_i = 1) = 1,\ i = 1, \dots, n$); it does, however, allow for a queue. They show for their model that the limiting probabilities are the same no matter which idle time ordering policy is employed.

2. Model and preliminaries

Arrivals come to an n server system in accordance with a Poisson process with rate λ. Each arrival has a vector of binary values $(x_1, \dots, x_n)$ with the interpretation that server i is eligible to serve that arrival if $x_i = 1$ and is ineligible if $x_i = 0, i = 1, \dots, n$. The binary vectors of successive arrivals are independent and identically distributed having the distribution of $(X_1, \dots, X_n)$, where $X_1, \dots, X_n$ are assumed to be exchangeable. An arrival that finds all of its eligible servers busy is lost; otherwise an arrival can be assigned to any one of its eligible servers. The time it takes server i to serve a customer has a general distribution $G_i, i = 1, \dots, n$.

Set $\beta_0 = 0$ and, for $k > 0$, let $\beta_k = \mathbb{P}(\sum_{i=1}^{k} X_i > 0)$; thus, by the exchangeability assumption, β_k is the probability that at least one of any specified set of k servers is eligible for the next job. Note that

$$\mathbb{P}\biggl(\sum_{i=1}^{k-1} X_i = 0,\ X_k = 1\biggr) = \beta_k - \beta_{k-1}, \qquad k \geq 1.$$

Let the 'idle servers vector' be 0 if there are currently no idle servers, or $i_1, \dots, i_k$ if there are currently k idle servers, with i_1 having been idle the longest, i_2 the second longest, and so on. Define an idle time ordering rule as one whose assignment decisions are based solely on the number of idle servers and the positions in the idle servers vector of those that are eligible to be assigned. Each idle time ordering rule results in a set of probabilities $\mathbb{P}_{j,k},\ j \leq k \leq n$, where $\mathbb{P}_{j,k}$ is the probability that a job arrival when the idle servers vector is $i_1, \dots, i_k$ will be assigned to server i_j. Note that $\sum_{j=1}^{k} \mathbb{P}_{j,k} = \beta_k$.

Examples. (a) If the idle time ordering rule in use is to give an incoming job to the idle eligible server that has been idle the longest, then $\mathbb{P}_{j,k} = \beta_j - \beta_{j-1}$.

(b) If the idle time ordering rule in use is to give an incoming job to the idle eligible server that has been idle the shortest, then $\mathbb{P}_{j,k} = \beta_{k+1-j} - \beta_{k-j}$.

(c) If the idle time ordering rule in use is to give an incoming job to a randomly chosen idle eligible server, then $\mathbb{P}_{j,k} = \beta_k / k$.

https://doi.org/10.1239/jap/1429282621 Published online by Cambridge University Press

3. General Erlang random variables and the method of stages

A random variable X is called a general Erlang $\mathrm{GE}(N, \mu)$ random variable if it can be expressed as

$$X = \sum_{i=1}^{N} W_i,$$

where W_i, $i \geq 1$, are independent exponential random variables with rate μ, and N is a positive integer valued random variable that is independent of W_i, $i \geq 1$.

Let G be the distribution function of a $\mathrm{GE}(N, \mu)$ random variable X, and let G_{e} be the equilibrium distribution of G. That is,

$$G_{\mathrm{e}}(x) = \frac{\mu}{\mathbb{E}[N]} \int_0^x (1 - G(y))\,\mathrm{d}y,$$

where $\mathbb{E}$ signifies expectation. This leads to the following lemma.

Lemma 1. *Let G_{e} be the distribution function of a* $\mathrm{GE}(N_{\mathrm{e}}, \mu)$ *random variable, where*

$$\mathbb{P}(N_{\mathrm{e}} = j) = \frac{\mathbb{P}(N \geq j)}{\mathbb{E}[N]}.$$

Proof. Let G_{e} be the limiting distribution of the excess of a renewal process having interarrival distribution G. Because G is a $\mathrm{GE}(N, \mu)$ distribution, we can interpret G_{e} as the limiting distribution of the time until the next visit to state 1 of a continuous-time Markov chain which spends an exponential time with rate μ in each state, and which when leaving state i goes to state j with probability $\mathbb{P}_{i,j}$, where

$$\mathbb{P}_{i,i-1} = 1, \quad i > 1, \qquad \mathbb{P}_{1,j} = \mathbb{P}(N = j), \quad j \geq 1.$$

It is well known (and quite easy to verify) that the limiting probabilities of the preceding embedded Markov chain are $\pi_i = \mathbb{P}(N \geq i)/\mathbb{E}[N]$. Because the time spent in each state has the same distribution, π_i, $i \geq 1$, is also the limiting probability distribution for the state of the continuous-time Markov chain. But if the state of the continuous-time chain is i then the time until the next visit to state 1 is distributed as the sum of i independent exponentials with rate μ, which proves the lemma.

For every nonnegative random variable Z there is a sequence of general Erlang variables that converges in distribution to Z; see, for instance, [13]. Because of this, our approach will be to first assume that all service distributions are general Erlang. By imagining that a service time of server i consists of N_i exponential stages, with the times of these stages being independent exponentials with rate μ_i, we are able to analyze the resulting model as a continuous-time Markov chain. This method of approximating any arbitrary distributed positive random variable by a general Erlang distribution and then analyzing the resultant model as a Markov chain is called the method of stages and has been used to analyze various models with general distributions (see, for instance, [10]–[12] and [7]).

To show that the results we obtain under the assumption that all service distributions are general Erlang remain valid when the service distributions are arbitrary requires that our quantities of interest are continuous functions of the service distributions. Barbour [2] and Whitt [13] have shown that such a continuity exists.

https://doi.org/10.1239/jap/1429282621 Published online by Cambridge University Press

4. Stationary probabilities for no-memory rules with general service time distribution

To start the process, we will suppose that the service distribution of server i is general Erlang $\mathrm{GE}(N_i, \mu_i)$, $i = 1, \dots, n$ and we will analyze the model as a continuous-time Markov chain. To do so, we define the state vector as $(0\colon \boldsymbol{r})$ with $\boldsymbol{r} = (r_1, \dots, r_n)$, if there are currently no idle servers and server i has r_i remaining exponential stages with rate μ_i in order to complete its service; or as $(i_1, \dots, i_k\colon \boldsymbol{r})$, if $i_1, \dots, i_k$ are the idle servers, with i_1 having been idle the longest, i_2 the second longest, and so on, and each server i has r_i exponential stages to complete, where $r_i = 0$ for all $i \in \{i_1, \dots, i_k\}$.

Proposition 1. *For general Erlang service times,* $\mathrm{GE}(N_i, \mu_i)$ $i = 1, \dots, n$, *where* N_i *is a random variable with* $p_i(s) = \mathbb{P}(N_i = s)$, *all idle time ordering policies have the same stationary probabilities. Namely,*

$$\mathbb{P}(i_1, \dots, i_k\colon \boldsymbol{r}) = \frac{\mu_{i_1} \cdots \mu_{i_k}}{\lambda^k \beta_1 \cdots \beta_k} \mathbb{P}(0\colon \mathbf{1}) \prod_{m \notin \{i_1, \dots, i_k\}} \mathbb{P}(N_m \geq r_m),$$

where $\mathbb{P}(0\colon \mathbf{1})$ *is such that*

$$\mathbb{P}(0\colon \mathbf{1})\Big(1 + \sum_{(i_1, \dots, i_k, \boldsymbol{r})} \frac{\mu_{i_1} \cdots \mu_{i_k}}{\lambda^k \beta_1 \cdots \beta_k} \prod_{m \notin \{i_1, \dots, i_k\}} \mathbb{P}(N_m \geq r_m)\Big) = 1.$$

Proof. Suppose that an arbitrary idle time ordering policy is being used, and let $\mathbb{P}_{j,k}$ be the probability, under that policy, that an arrival goes to server i_j when the state is $(i_1, \dots, i_k\colon \boldsymbol{r})$. For states

$$\begin{aligned}
\boldsymbol{x} &= (i_1, \dots, i_k\colon r_1, \dots, r_n),\\
\boldsymbol{x}^o &= (i_1, \dots, i_k\colon r_1, \dots, r_j - 1, \dots, r_n),\\
\boldsymbol{x}^+ &= (i_1, \dots, i_k, i_{k+1}\colon r_1, \dots, r_{i_{k+1}-1}, 0, r_{i_{k+1}+1}, \dots, r_n),\\
\boldsymbol{x}^- &= (i_1, \dots, i_{j-1}, i_{j+1}, \dots, i_k\colon r_1, \dots, r_{i_j-1}, s, r_{i_j+1}, \dots, r_n),
\end{aligned}$$

the infinitesimal rates of the resultant continuous-time Markov chain are

$$\begin{aligned}
q_{\boldsymbol{x},\boldsymbol{x}^o} &= \mu_j, \quad \text{for } r_j > 1,\\
q_{\boldsymbol{x},\boldsymbol{x}^+} &= \mu_{i_{k+1}}, \quad \text{for } r_{i_{k+1}} = 1,\\
q_{\boldsymbol{x},\boldsymbol{x}^-} &= \lambda \mathbb{P}_{j,k} p_{i_j}(s) \quad \text{for } r_{i_j} = 0.
\end{aligned}$$

We now make the following conjecture about the reverse process.

(a) It is a queueing model with n servers all of whom are eligible to serve any arriving customer.

(b) The state is $\boldsymbol{x} = (i_1, \dots, i_k\colon r_1, \dots, r_n)$ if $(i_1, \dots, i_k)$ is the current ordered list of idle servers; r_j is the current stage of server j if that server is busy (meaning that $r_j - 1$ stages have already been completed), and $r_j = 0$ if j is idle.

(c) An arrival to server j begins in stage 1; and the time it takes server j to complete a stage is exponential with rate μ_j, $j = 1, \dots, n$.

(d) Upon completion of stage m, a customer at server j leaves the system with probability $\lambda_j(m) = p_j(m) / \sum_{k \geq m} p_j(k)$; otherwise it goes to stage $m+1$ with probability $\bar{\lambda}_j(m) = 1 - \lambda_j(m)$.

https://doi.org/10.1239/jap/1429282621 Published online by Cambridge University Press

(e) The arrival rate of customers when the ordered list of idle servers is $i_1, \ldots, i_k$ is $\lambda\beta_k$, and the arriving customer is assigned to server i_k.

(f) If server r becomes idle when the ordered list of idle servers is $i_1, \ldots, i_k$ then the new ordered list of idle servers becomes $i_1, \ldots, i_{j-1}, r, i_j, \ldots, i_k$ with probability $\mathbb{P}_{j,k+1}/\beta_{k+1}$.

Under our conjecture, with $\boldsymbol{x}, \boldsymbol{x}^o, \boldsymbol{x}^+$, and $\boldsymbol{x}^-$ as previously defined, the infinitesimal rates of the reversed chain are

$$q^*_{\boldsymbol{x}^o,\boldsymbol{x}} = \mu_j \, \bar{\lambda}_j(r_j - 1), \qquad q^*_{\boldsymbol{x}^+,\boldsymbol{x}} = \lambda\beta_{k+1}, \qquad q^*_{\boldsymbol{x}^-,\boldsymbol{x}} = \mu_{i_j}\lambda_{i_j}(s)\frac{\mathbb{P}_{j,k}}{\beta_k}.$$

Because it is clear that when in state $(i_1, \ldots, i_k \colon \boldsymbol{r})$ the rates at which the forward and the conjectured reverse process leave that state are both equal to $\lambda\beta_k + \sum_{i\notin\{i_1,\ldots,i_k\}} \mu_i$, it follows from Theorem 1.13 of [8] that the conjecture will be verified if we can find probabilities $\mathbb{P}(\boldsymbol{x})$ such that $\sum_{\boldsymbol{x}} \mathbb{P}(\boldsymbol{x}) = 1$, and

$$\begin{aligned} \mathbb{P}(\boldsymbol{x})q_{\boldsymbol{x},\boldsymbol{x}^o} &= \mathbb{P}(\boldsymbol{x}^o)q^*_{\boldsymbol{x}^o,\boldsymbol{x}} && \text{for } r_j > 1, \\ \mathbb{P}(\boldsymbol{x})q_{\boldsymbol{x},\boldsymbol{x}^+} &= \mathbb{P}(\boldsymbol{x}^+)q^*_{\boldsymbol{x}^+,\boldsymbol{x}} && \text{for } r_{i_{k+1}} = 1, \\ \mathbb{P}(\boldsymbol{x})q_{\boldsymbol{x},\boldsymbol{x}^-} &= \mathbb{P}(\boldsymbol{x}^-)q^*_{\boldsymbol{x}^-,\boldsymbol{x}} && \text{for } r_{i_j} = 0. \end{aligned}$$

Thus, we must find probabilities that satisfy

$$\mathbb{P}(\boldsymbol{x})\mu_j = \mathbb{P}(\boldsymbol{x}^o)\bar{\lambda}_j(r_j - 1)\mu_j \quad \text{for } r_j > 1, \tag{1}$$

$$\mathbb{P}(\boldsymbol{x})\mu_{i_{k+1}} = \mathbb{P}(\boldsymbol{x}^+)\lambda\beta_{k+1} \quad \text{for } r_{i_{k+1}} = 1, \tag{2}$$

$$\mathbb{P}(\boldsymbol{x})\lambda\mathbb{P}_{j,k}p_{i_j}(s) = \mathbb{P}(\boldsymbol{x}^-)\mu_{i_j}\lambda_{i_j}(s)\frac{\mathbb{P}_{j,k}}{\beta_k} \quad \text{for } r_{i_j} = 0. \tag{3}$$

Using the fact that $\mathbb{P}(N_j \geq m)\bar{\lambda}_j(m) = \mathbb{P}(N_j \geq m + 1)$ and, analogously, that $\mathbb{P}(N_j \geq m)\lambda_j(m) = p_j(m)$, it is easily checked that

$$\mathbb{P}(i_1, \ldots, i_k \colon \boldsymbol{r}) = \frac{\mu_{i_1} \cdots \mu_{i_k}}{\lambda^k \beta_1 \cdots \beta_k}\mathbb{P}(0, \mathbf{1}) \prod_{m\notin\{i_1,\ldots,i_k\}} \mathbb{P}(N_m \geq r_m)$$

satisfy (1), (2), and (3). Hence, with $\mathbb{P}(0, \mathbf{1})$ chosen to make the probabilities sum to 1, the proposition is proven.

Theorem 1. *Suppose that the service distributions $G_1, \ldots, G_n$ are arbitrary, and let $\mathbb{E}[S_j]$ be the mean of the distribution G_j. If $\mathcal{I}$ is the set of idle servers in steady state then*

$$\mathbb{P}(\mathcal{I} = \{i_1, \ldots, i_k\}) = k! \, \frac{1}{\lambda^k \beta_1 \cdots \beta_k \mathbb{E}(S_{i_1}) \cdots \mathbb{E}(S_{i_k})}\mathbb{P}(0),$$

where $\mathbb{P}(0)$ is the probability that all servers are busy. Furthermore, given that $\mathcal{I} = \{i_1, \ldots, i_k\}$:

(a) *all $k!$ possible orderings of the idle servers are equally likely;*

(b) *the remaining service times of the busy servers are independent and are distributed according to their respective equilibrium service distributions;*

(c) *the amounts of service time already provided on their current customers by the busy servers are independent and are distributed according to their respective equilibrium service distributions.*

https://doi.org/10.1239/jap/1429282621 Published online by Cambridge University Press

Proof. Suppose that the service distribution of server i is general Erlang $\mathrm{GE}(N_i, \mu_i)$, $i = 1, \dots, n$. Let $\mathbb{P}(i_1, \dots, i_k)$ be the steady state probability that $i_1, \dots, i_k$ is the idle servers vector. Using Proposition 1 and summing $\mathbb{P}(i_1, \dots, i_k \colon \boldsymbol{r})$ over all the consistent vectors $\boldsymbol{r}$ (that is, all $\boldsymbol{r}$ such that $r_j = 0$, $j \in \{i_1, \dots, i_k\}$) yields

$$\mathbb{P}(i_1, \dots, i_k) = \frac{\mu_{i_1} \cdots \mu_{i_k}}{\lambda^k \beta_1 \cdots \beta_k} \mathbb{P}(0 \colon \mathbf{1}) \sum_{\boldsymbol{r}} \prod_{m \notin \{i_1, \dots, i_k\}} \mathbb{P}(N_m \geq r_m).$$

Letting $\{b_1 \dots, b_{n-k}\}$ be the complement of the set $\{i_1, \dots, i_k\}$, the preceding equation yields

$$\begin{aligned}\mathbb{P}(i_1, \dots, i_k) &= \frac{\mu_{i_1} \cdots \mu_{i_k}}{\lambda^k \beta_1 \cdots \beta_k} \mathbb{P}(0 \colon \mathbf{1}) \sum_{r_{b_1}} \cdots \sum_{r_{b_{n-k}}} \mathbb{P}(N_{b_1} \geq r_{b_1}) \cdots \mathbb{P}(N_{b_{n-k}} \geq r_{b_{n-k}}) \\ &= \frac{\mu_{i_1} \cdots \mu_{i_k}}{\lambda^k \beta_1 \cdots \beta_k} \mathbb{P}(0 \colon \mathbf{1}) \prod_{m \notin \{i_1, \dots, i_k\}} \mathbb{E}(N_m). \qquad (4)\end{aligned}$$

With $\mathbb{P}(0)$ equal to the probability that all servers are busy, from Proposition 1 it follows that

$$\mathbb{P}(0) = \sum_{\boldsymbol{r}} \mathbb{P}(0 \colon \boldsymbol{r}) = \sum_{\boldsymbol{r}} \mathbb{P}(0 \colon \mathbf{1}) \prod_{i=1}^{n} \mathbb{P}(N_i \geq r_i) = \mathbb{P}(0 \colon \mathbf{1}) \prod_{i=1}^{n} \mathbb{E}(N_i).$$

Using the preceding equation, along with $\mathbb{E}(S_i) = \mathbb{E}(N_i)/\mu_i$, $i = 1, \dots, n$, allows us to rewrite (4) as

$$\mathbb{P}(i_1, \dots, i_k) = \frac{1}{\lambda^k \beta_1 \cdots \beta_k \mathbb{E}(S_{i_1}) \cdots \mathbb{E}(S_{i_k})} \mathbb{P}(0),$$

which yields

$$\mathbb{P}(\mathcal{I} = \{i_1, \dots, i_k\}) = k! \, \frac{1}{\lambda^k \beta_1 \cdots \beta_k \mathbb{E}(S_{i_1}) \cdots \mathbb{E}(S_{i_k})} \mathbb{P}(0),$$

as well as showing that, conditional on $\mathcal{I} = \{i_1, \dots, i_k\}$, all $k!$ possible orderings of idle servers are equally likely.

Moreover, it follows from Proposition 1 and (4) that

$$\frac{\mathbb{P}(i_1, \dots, i_k; \boldsymbol{r})}{\mathbb{P}(i_1, \dots, i_k)} = \prod_{m \notin \{i_1, \dots, i_k\}} \frac{\mathbb{P}(N_m \geq r_m)}{\mathbb{E}(N_m)},$$

which, using Lemma 1, proves that conditional on the set of busy servers their remaining service times are independent and are distributed according to their respective equilibrium service distributions. In addition, because the reverse chain has the same stationary probabilities as does the forward chain, and as the interpretation of r_i for the reverse chain is that server i is currently at stage r_i, part (c) also follows. Hence, the theorem is proven when all service distributions are of general Erlang type. Because any service distribution is the limit of a sequence of general Erlang distributions, the approach of Barbour [2] can now be used to establish the necessary continuity. Thus, the theorem is proven for arbitrary service distributions.

https://doi.org/10.1239/jap/1429282621 Published online by Cambridge University Press

5. Utilizing a Markov chain Monte Carlo simulation method

When n is large the determination of the constant $\mathbb{P}(0)$ is computationally intractable. Indeed, even if it were known, the derivations of other quantities of interest, such as average waiting time in the system, rate at which customers are lost, etc., remain computationally intractable. However, these quantities can be determined by using the Gibbs sampler Markov chain Monte Carlo method to generate a Markov chain whose limiting distribution is the stationary distribution of the set of idle servers. That is, interpreting Y_i as the indicator of whether server i is idle, we want to generate a Markov chain whose stationary distribution is

$$p(x_1, \dots, x_n) = \mathbb{P}(Y_i = x_i,\ i = 1, \dots, n) = C\frac{k!}{\lambda^k \beta_1 \cdots \beta_k \prod_{i=1}^n x_i \mathbb{E}[S_i]},$$

where $x_i = 0, 1,\ k = \sum_{i=1}^n x_i$.

When the current state of the Markov chain is $\boldsymbol{x} = (x_1, \dots, x_n)$, the Gibbs sampler method chooses a coordinate that is equally likely to be any of $1, \dots, n$. If coordinate j is chosen then the next state will be $(x_1, \dots, x_{j-1}, 0, x_{j+1}, \dots, x_n)$ with probability

$$\begin{aligned}\alpha &= \frac{p(x_1, \dots, x_{j-1}, 0, x_{j+1}, \dots, x_n)}{p(x_1, \dots, x_{j-1}, 1, x_{j+1}, \dots, x_n) + p(x_1, \dots, x_{j-1}, 0, x_{j+1}, \dots, x_n)}\\ &= \frac{\lambda\mathbb{E}(S_j)\beta_{r+1}}{r + 1 + \lambda\mathbb{E}(S_j)\beta_{r+1}},\end{aligned}$$

where $r = \sum_{i \neq j} x_i$, or it will be $(x_1, \dots, x_{j-1}, 1, x_{j+1}, \dots, x_n)$ with probability $1 - \alpha$.

The stationary distribution of the successive values of a Markov chain generated by the preceding equation is the limiting distribution of the set of idle servers. Consequently, we can approximate the steady state probability that there are exactly k idle servers, call it $\mathbb{P}(k)$, by the proportion of states $(x_1, \dots, x_n)$ such that $\sum_{i=1}^n x_i = k$. We can then use our estimates of $\mathbb{P}(k)$, $k = 0, \dots, n$, to estimate $\sum_{k=1}^n \mathbb{P}(k)\beta_k$, equal to the proportion of arrivals that enter the system. Similarly, we can estimate the proportion of arrivals that are served by server i by letting $\pi(i\colon k)$ be the proportion of states of the chain for which $x_i = 1$, $\sum_{j=1}^n x_j = k$ and then using the estimate $\sum_{k=1}^n \pi(i\colon k)\beta_k/k$, where the preceding uses the fact that conditional on the set of idle servers all orderings are equally likely, and so, given that an arrival is eligible, each server in this set is equally likely to be the one used.

Example 1. In Table 1 we compare the efficiency of the Gibbs sampler method in finding the multiplicative constant with a discrete event Monte Carlo simulation. We simulated a model with five servers with deterministic service times, $D_i = 1/i$, $i = 1, \dots, 5$. The arrival process is a Poisson process with rate $\lambda = 15$. We have also assumed that the component of the eligibility vectors are independent Bernoulli random variables with mean $\frac{1}{2}$.

The results are based on 500 runs where each run has a length equal to three seconds. The per run estimator in the discrete event simulation is the proportion of the run simulated time that all servers are busy; the per run estimator in the Gibbs sampler approach is the proportion of states having all of its components equal to 0. The actual value of $\mathbb{P}(0)$ was analytically determined and, for each run, we computed A, the absolute value of the difference between the run estimator and $\mathbb{P}(0)$. In Table 1 we list both the sample mean and the sample variance of the observed A values for the two methods. The estimators of $\mathbb{P}(0)$ are the average of the estimators over all 500 runs.

https://doi.org/10.1239/jap/1429282621 Published online by Cambridge University Press

TABLE 1: Estimate for $\mathbb{P}(0)$. Exact value of $\mathbb{P}(0) = 0.151\,49$.

Estimates of	Gibbs sampler	Raw simulation
$\mathbb{P}(0)$	0.147 79	0.144 40
$\mathbb{E}(A)$	0.028 13	0.032 10
$\mathrm{var}(A)$	0.000 432 5	0.000 891 3

References

[1] ADAN, I. AND WEISS, G. (2012). A loss system with skill-based servers under assign to longest idle server policy. *Prob. Eng. Inf.. Sci.* **26,** 307–321.

[2] BARBOUR, A. D. (1976). Networks of queues and the method of stages. *Adv. Appl. Prob.* **8,** 584–591.

[3] COOPER, R. B. AND S. PALAKURTHI. (1989). Heterogeneous-server loss systems with ordered entry: an anomaly. *Operat. Res. Lett.* **8,** 347–349.

[4] FAKINOS, D. (1980). The *M/G/k* blocking system with heterogeneous servers. *J. Operat. Res. Soc.* **31,** 919–927.

[5] FAKINOS, D. (1982). The generalized *M/G/k* blocking system with heterogeneous servers. *J. Operat. Res. Soc.* **33,** 801–809.

[6] GOPALAKRISHNAN, R., DOROUDI, S., WARD, A. R., AND WIERMAN, A. (2015). Routing and Staffing when Servers are Strategic. Working paper.

[7] KELLY, F. P. (1976). Networks of queues. *Adv. Appl. Prob.* **8,** 416–432.

[8] KELLY, F. P. (1979). *Reversibility and Stochastic Networks*. John Wiley, Chichester.

[9] ROSS, S. M. (2014). Optimal server selection in a queueing loss model with heterogeneous exponential servers, discriminating arrivals, and arbitrary arrival times. *J. Appl. Prob.* **51,** 880–884.

[10] SCHASSBERGER, R. (1976). On the equilibrium distribution of a class of finite-state generalized semi-Markov processes. *Math. Operat. Res.* **1,** 395–406.

[11] SCHASSBERGER, R. (1977). Insensitivity of steady-state distributions of generalized semi-Markov processes. I. *Ann. Prob.* **5,** 87–99.

[12] SCHASSBERGER, R. (1978). Insensitivity of steady-state distributions of generalized semi-Markov processes. II. *Ann. Prob.* **6,** 85–93.

[13] WHITT, W. (1974). The continuity of queues. *Adv. Appl. Prob.* **6,** 175–183.

https://doi.org/10.1239/jap/1429282621 Published online by Cambridge University Press